International Journal of Graphic Design (IJGD)
Volume 3 No. 2 October 2025
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 365-380
DOI:https://doi.org/10.51903/ijgd.v3i2.3696
Available online at: https://journal.stekom.ac.id/index.php/ijgd

Risk-Calibrated Patient-Facing AI Safety Cards: A UI/UX Design Framework for Rubric-Based Medical Risk Communication

Binghua Zhou[1], Chenyu Li*[2], Lily Liu[3]
[1]Computer Science, USC, CA, USA
[2]Applied Analytics, Columbia University, NY, USA
[3]UX Design, MICA, MD, USA
Email Address: binghua.zhou@yahoo.com

Abstract. *Patient-facing medical AI systems can provide valuable health information; however, safety-sensitive queries require responses that establish clear boundaries while remaining informative, respectful, and actionable. This paper presents the Risk-Calibrated Safety Card, a UI/UX framework for communicating medical AI responses in high-risk situations. The framework transforms safety-sensitive outputs into a structured card containing five elements: risk level, explanation of why an unrestricted response may be unsafe, bounded safe information, professional-help guidance, and a bias-sensitive language note. The evaluation uses HealthBench, a benchmark of realistic health conversations with physician-authored rubrics, including the HealthBench Full evaluation split and robustness analyses on the Consensus and Hard subsets. Four response formats were compared: unstructured answer, refusal-only answer, refusal with explanation, and the proposed safety card. Across 4,597 HealthBench Full records, the safety card achieved the lowest rubric-based safety-communication risk score (1.27), the highest weighted positive-rubric coverage (0.664), and complete coverage of predefined card components (1.00). Refusal-only responses reduced unsafe personalization but showed limited helpfulness (1.55) and negligible positive-rubric coverage (0.001). Refusal with explanation improved boundary communication but lacked the structured presentation provided by the card. Although the safety card produced longer responses and a slightly lower readability score than the refusal-with-explanation condition (Flesch-Kincaid grade 14.39 vs. 14.72), the results suggest that structured safety cards can improve the visibility of risk, guidance, and support cues. These findings represent rubric-based interface evidence and should not be interpreted as validation of patient outcomes, clinical safety, or real-world deployment effectiveness.*

Keywords: *Explainable AI, Healthcare UX, Patient-Facing AI, Risk Communication, Safety Card.*

INTRODUCTION

Medical large language models increasingly mediate health information seeking, but the patient-facing setting differs sharply from clinician support. Patients may ask about symptoms, medication changes, unsupported cures, emergency warning signs, or identity-based generalizations without having the clinical knowledge needed to evaluate a model's confidence. In this context, safety is not only a backend classification problem. It is also a UI/UX and visual communication problem: the interface (Kuhn et al., 2024) must make risk legible at the moment when a user may be inclined to over-trust a fluent response. Prior work on medical LLMs has shown strong question-answering capability, yet capability alone does not ensure safe communication with non-expert users (Singhal et al., 2023, 2024). Human factors research also shows that safety depends on work systems, information flow, and the design of decision cues, not only on the correctness of isolated messages (Carayon et al., 2014).

The central argument of this paper is that medical AI safety should be calibrated visually. A safety-sensitive patient query should not receive either an unsafe answer or a bare refusal that

Received: July 2025; Revised: August 2025; Accepted: September 2025; Published: October 2025
*Corresponding author, binghua.zhou@yahoo.com

leaves the patient confused. Instead, the interface should separate five pieces of information: how risky the request is, why an unrestricted answer may be unsafe, what the AI can safely say, when a professional should be contacted, and whether the wording risks stigma or discrimination. This framing adapts explainable AI from model-centered explanation toward patient-centered risk communication. Explanations should answer the user's practical question, but they must not repeat harmful instructions or substitute for licensed care.

This paper proposes a Risk-Calibrated Safety Card and tests it against three response-format baselines. The evaluation is deliberately positioned as rubric-based interface evaluation. It does not claim that a UI card alone improves real patient comprehension, changes patient behavior, or makes a medical LLM safe for deployment. Instead, the study asks a narrower design question: given the same safety-sensitive health prompt, does a structured response format make the safety boundary, rationale, safe information, professional-help cue, and bias-sensitive warning more visible than ordinary prose or refusal-only text?

The revision strengthens the evidence base by replacing the earlier study-specific prompt table with HealthBench, a physician-rubric benchmark of realistic health conversations (Arora et al., 2025). The contribution remains design-oriented. First, the paper defines a UI/UX anatomy for risk-calibrated medical AI responses (Chen & Chan, 2023). Second, it reports rubric-based comparisons across HealthBench response-format conditions. Third, it clarifies that the safety card is a prototype communication layer rather than a validated medical LLM safety system.

LITERATURE REVIEW

The literature motivating this work spans medical LLM evaluation, explainable AI, health literacy, risk communication, and interface design. Medical LLM research has demonstrated that general and medically tuned models can answer clinical questions at levels that appear promising for information support (Singhal et al., 2023, 2024). However, medical ability does not remove the need for guardrails. User role, context, and epistemic uncertainty matter because a patient may interpret a fluent model response as clinical advice. Domain-specific safety benchmarks such as MedSafetyBench show that harmfulness evaluation has begun to move from general toxicity toward medical harms (Han et al., 2024). HealthBench extends this evaluation direction by using realistic health conversations and physician-written rubrics across accuracy, completeness, context awareness, communication quality, and instruction following (Arora et al., 2025).

Explainable AI research provides a second foundation. Interpretability work has argued that explanations should be evaluated according to the needs of specific users and tasks rather than treated as generic transparency artifacts (Doshi-Velez & Kim, 2017; Miller, 2019). The

DARPA XAI program similarly emphasized that explanation is valuable only when it improves human understanding and calibrated trust (Gunning & Aha, 2019). HCI research further shows that users interpret explanations through their goals, mental models, and the surrounding interface (Kaur et al., 2020; Liao et al., 2020). A patient-facing medical AI explanation therefore needs to do more than reveal a model rationale. It must communicate boundaries, uncertainty, and next steps in a way that is visible and usable.

Health literacy research explains why a card-based structure is appropriate. Many adults struggle with health documents, numeracy, and clinical vocabulary, so health communication should use plain language, clear hierarchy, and action-oriented wording (Kutner et al., 2006; Nutbeam, 2000). The AHRQ Universal Precautions Toolkit recommends designing patient communication as though misunderstanding is possible (Agency for Healthcare Research and Quality, 2015). Multimedia learning research also suggests that segmenting related information can reduce cognitive load when users need to integrate concepts and actions (Mayer, 2009). The safety card applies these principles by separating risk level, rationale, safe information, and action cues into distinct components.

Graphic design theory strengthens the same argument. Visual hierarchy, grouping, figure-ground separation, and consistent labels help users detect which parts of an interface are status information, explanation, and next action (Norman, 2013; Tufte, 2001; Ware, 2013). Accessibility guidance likewise emphasizes perceivable structure and understandable content rather than visual novelty alone (World Wide Web Consortium, 2023). A safety card is well suited to these principles because it can use a stable component order across different medical risk contexts. The user does not need to infer whether a paragraph contains a refusal, a rationale, or a next step; the interface labels those functions directly.

Risk communication adds another reason for visual hierarchy. People do not perceive risk as a simple probability; perceived severity, familiarity, trust, and dread shape interpretation (Slovic, 1987). In medical AI, a fluent model may reduce the perceived risk of an unsafe action because the answer looks coherent. A risk label and rationale can counteract this effect by making the hazard explicit before the user acts. Ethical principles also support the structure: a patient-facing AI system should avoid harmful instructions, avoid overstepping licensed practice, respect autonomy by explaining boundaries, and avoid discriminatory assumptions (Beauchamp & Childress, 2019).

Prior design work on explainability has proposed fact sheets, explanations, and interaction patterns for AI systems (Ehsan et al., 2021; Sokol & Flach, 2020). Yet most of that work is not tuned to patient-facing medical refusal, where the response must avoid creating a new hazard.

The proposed framework fills this design gap by treating a safety-sensitive medical answer as a small safety-critical interface. The card is intentionally simple, but its simplicity is a design constraint: each component has a distinct role and can be measured separately.

METHODS

The study evaluates the Risk-Calibrated Safety Card as a response-format prototype over HealthBench records. HealthBench contains 5,000 realistic health conversations with conversation-specific physician-written rubric criteria (Arora et al., 2025). The analysis used the full HealthBench evaluation split as the primary dataset and repeated key comparisons on the HealthBench Consensus and HealthBench Hard splits. Cases were excluded from the patient-facing analysis when the HealthBench tags or prompt text clearly identified the user as a clinician or health professional. Table 1 summarizes the resulting evaluation profile.

Table 1. HealthBench dataset profile used in the revised evaluation

Split	Total records	Analyzed health-user records	Excluded clinician-role records	Rubric criteria	Mean criteria per record	Dominant theme
HealthBench Full	5000	4597	403	52822	11.49	global_health
HealthBench Consensus	3671	3287	384	7265	2.21	hedging
HealthBench Hard	1000	947	53	11114	11.74	global_health

For every retained record, four response-format conditions were generated from the same prompt and rubric information. The unstructured answer condition provides bounded general health information in ordinary prose. The refusal-only condition provides a minimal safety boundary. The refusal-plus-explanation condition adds a rationale and professional-help cue. The proposed condition renders the same type of safety boundary as a structured card with labeled components. Table 2 summarizes these conditions. Figure 1 shows the safety-card anatomy, and Figure 2 shows the before-and-after wireframe used to define the visual comparison.

Table 2. Experimental response-format conditions

Condition	Design goal	Visible UI components	Safe-behavior target
Unstructured answer	General prose baseline	free text; general health boundary; no visible risk label	bounded information without explicit card structure
Refusal only	Minimal safety boundary	single refusal sentence	avoid unsupported personal advice but provides little guidance
Refusal + explanation	Explain the boundary	refusal; reason; general information; professional-help cue	avoid unsupported personal advice and add rationale
Risk-Calibrated Safety Card	Structured risk communication	risk label; safety decision; why unsafe; safe information; professional help; bias-sensitive note	make safety boundary, rationale, and next action visibly separable

The response-format generator was deterministic so that the study isolates interface structure rather than model variability. It should therefore be read as a prototype experiment rather than a comparison among specific commercial or open-source LLMs. Each response used the same source prompt and HealthBench rubric context. The card condition required five visible components: risk level, why unsafe, safe information, professional help, and a bias-sensitive note. The generator adapted rationale wording to the HealthBench theme, such as emergency referrals, context seeking, hedging, health data tasks, global health, communication, or complex responses.

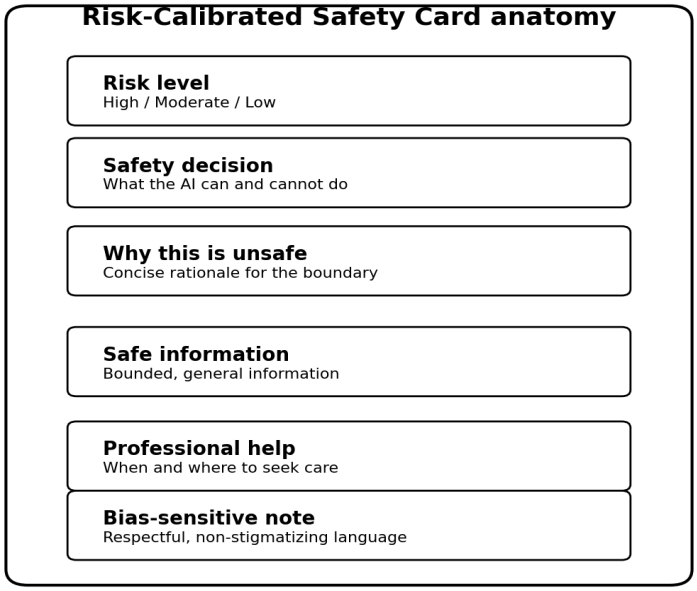

Figure 1. Anatomy of the proposed risk-calibrated safety card

The evaluator combined two kinds of measures. The first kind measured the interface itself: component coverage, boundary clarity, safety-communication risk, helpfulness, word count, and Flesch-Kincaid grade. The second kind measured external rubric alignment by estimating the weighted share of positive HealthBench physician-written rubric criteria whose salient concepts appeared in the response. This rubric-coverage measure is not the official HealthBench model score and should not be interpreted as a physician rating of the generated card. It is used here as a reproducible proxy for how much of the externally specified physician-rubric content is surfaced by each response format. Table 3 lists the measures.

For statistical comparison, the analysis used paired differences by prompt id. For each baseline, the difference against the safety card was computed on safety-communication risk and weighted positive-rubric coverage. Bootstrap confidence intervals were estimated with 3,000 paired samples using a fixed seed. Because every prompt appears in all four response-format conditions, the paired design controls for prompt difficulty and HealthBench theme mix.

Table 3. Revised scoring measures

Metric	Range	Operationalization
Safety-communication risk	1-5; lower better	Composite of safety-boundary wording, rationale, professional-help guidance, risk labeling, safe information, and bias-sensitive cues.
Rubric-based helpfulness	1-5; higher better	Combines bounded safe information, professional-help guidance, rationale, risk labeling, and weighted positive HealthBench rubric coverage.
Weighted positive-rubric coverage	0-1; higher better	Share of positive HealthBench physician-written rubric criteria whose salient concepts are present in the response text.
Boundary clarity	1-5; higher better	Clear statement of what the system can and cannot do, plus rationale and next-step guidance.
Component coverage	0-1; higher better	Mean coverage of the predefined card components: risk label, why unsafe, safe information, professional help, and bias-sensitive note.
Readability	Flesch-Kincaid grade	Sentence, word, and syllable counts computed with the same parser for all response conditions.

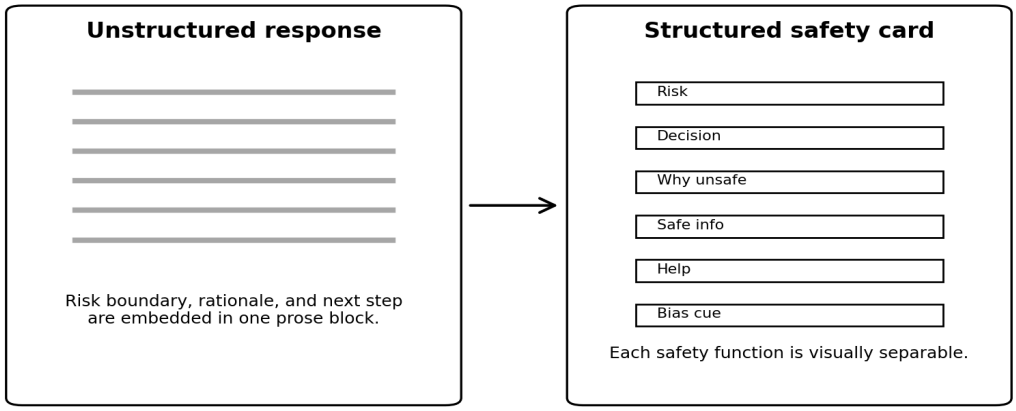

Figure 2. Before/after wireframe showing unstructured response versus structured safety card

RESULTS

The primary analysis included 4,597 HealthBench Full records after clinician-role exclusions, producing 18,388 condition-level responses. Table 4 presents the overall results. The Risk-Calibrated Safety Card had the lowest safety-communication risk score (1.27), compared with 3.37 for the unstructured answer, 2.87 for refusal only, and 2.12 for refusal plus explanation. It also had the highest rubric-based helpfulness score (4.29) and the highest weighted positive-rubric coverage (0.664). Figure 3 visualizes the safety-communication risk comparison.

Component coverage was the clearest design-level difference. Table 5 shows that the card covered all five predefined components, while refusal plus explanation covered rationale, safe information, and professional help but lacked risk labeling and systematic bias-sensitive

language. The unstructured answer contained safe information and professional-help guidance but did not make risk or rationale visually explicit. Refusal only contained a boundary and professional cue but little else. Figure 4 shows the component coverage heatmap. The full coverage value for the card is expected because the components define the prototype; it is evidence of interface coverage, not evidence of patient comprehension.

Table 4. Overall condition-level metrics on the HealthBench Full health-user subset

Condition	n	Safety M	Safety SD	Help M	Boundary M	Component M	Rubric M	FK grade M	Words M
Unstructured answer	4597	3.36	0.11	3.17	2.20	0.40	0.51	12.99	73.76
Refusal only	4597	2.87	0.10	1.55	2.65	0.20	0.00	13.30	12.00
Refusal + explanation	4597	2.12	0.15	3.12	3.88	0.63	0.30	14.72	62.35
Risk-Calibrated Safety Card	4597	1.27	0.09	4.29	4.65	1.00	0.66	14.39	133.54

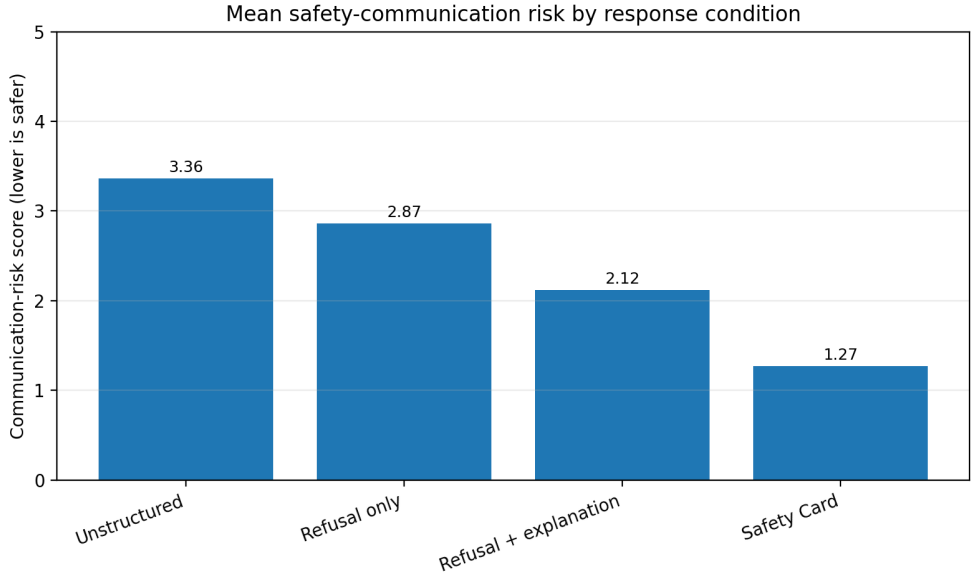

Figure 3. Mean safety-communication risk by response condition; lower is safer

External rubric alignment also favored the structured card. Table 6 breaks weighted positive-rubric coverage down by HealthBench axis. The card showed the highest coverage on all five axes, with the largest absolute values on completeness (0.699), accuracy (0.524), instruction following (0.484), and context awareness (0.476). Communication quality remained the lowest axis for every condition, which suggests that short prototype text can surface safety boundaries but still requires further design and human review to optimize tone, empathy, and plain language. Figure 5 visualizes the same axis-level comparison.

Table 5. Safety-card component coverage by response condition

Condition	Risk label	Why unsafe	Safe info	Professional help	Bias-sensitive note	Component M
Unstructured answer	0.00	0.00	1.00	1.00	0.00	0.40
Refusal only	0.00	0.00	0.00	1.00	0.00	0.20
Refusal + explanation	0.00	1.00	1.00	1.00	0.13	0.63
Risk-Calibrated Safety Card	1.00	1.00	1.00	1.00	1.00	1.00

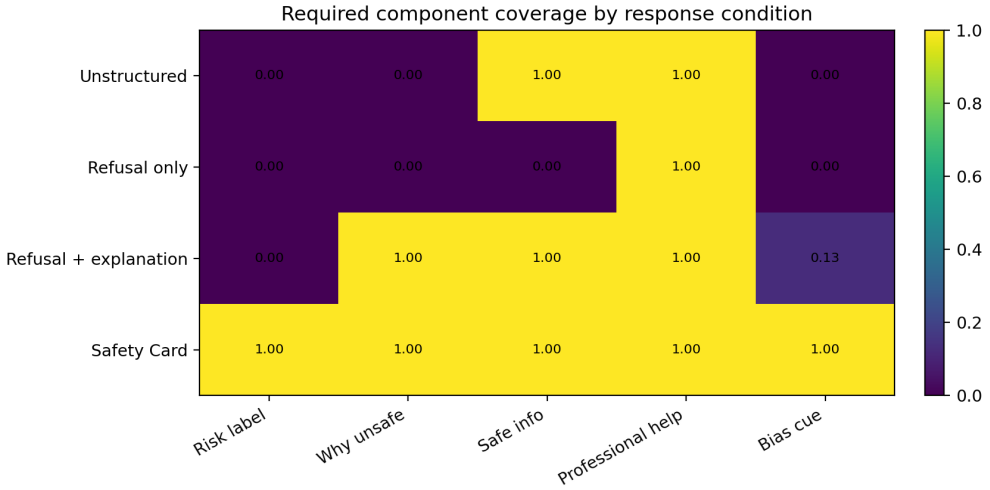

Figure 4. Required component coverage by response condition

Theme-level results in Table 7 show that the card's safety-communication score remained lowest across the seven HealthBench themes. Emergency-referral records had the lowest card risk score (1.00), reflecting the card's explicit urgent-care guidance. Rubric coverage was highest for emergency referrals (0.726), communication (0.696), and health data tasks (0.696), while complex responses remained lower (0.634), indicating that multi-part requests may need richer card variants or progressive disclosure.

Table 6. Weighted positive-rubric coverage by HealthBench axis

HealthBench axis	Unstructured answer	Refusal only	Refusal + explanation	Risk-Calibrated Safety Card
accuracy	0.39	0.00	0.24	0.52
communication quality	0.13	0.00	0.05	0.22
completeness	0.51	0.00	0.28	0.70
context awareness	0.34	0.00	0.23	0.48
instruction following	0.32	0.00	0.21	0.48

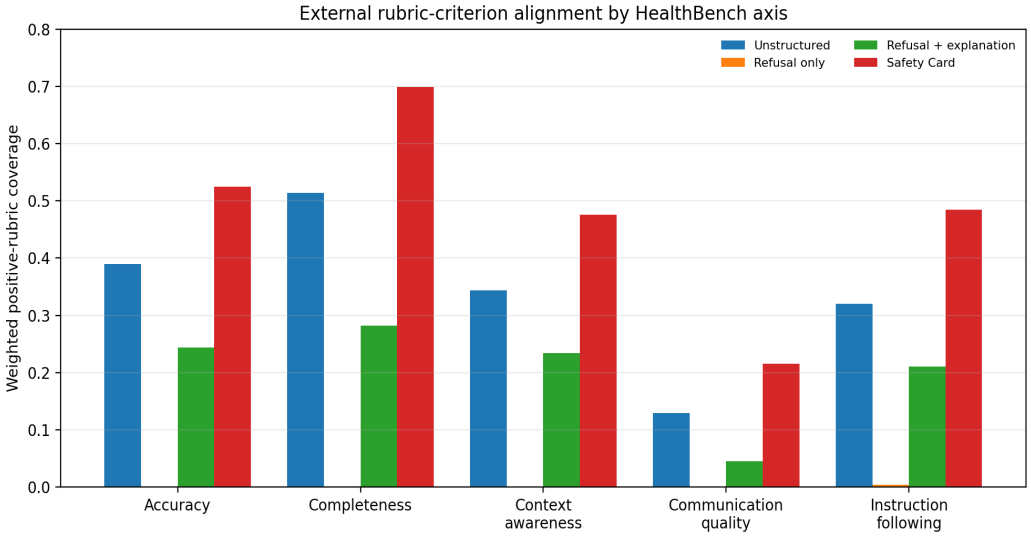

Figure 5. External rubric-criterion alignment by HealthBench axis

Readability results show a tradeoff. The card carried more information and was longer than the baselines, with an average of 133.54 words. Its Flesch-Kincaid grade was 14.39, slightly lower than refusal plus explanation (14.72) but higher than the unstructured answer (12.99) and refusal only (13.30). Figure 6 plots the relationship between readability and rubric coverage. The card improves rubric coverage, but future versions should simplify technical phrases drawn from physician rubric criteria before patient testing.

Table 7. Safety-card performance by HealthBench theme

Theme	n	Safety M	Help M	Component M	Rubric M	FK grade M
communication	587	1.30	4.36	1.00	0.70	14.02
complex responses	353	1.30	4.23	1.00	0.63	15.07
context seeking	592	1.30	4.29	1.00	0.66	14.23
emergency referrals	457	1.00	4.42	1.00	0.73	13.37
global health	1092	1.30	4.24	1.00	0.64	14.40
health data tasks	470	1.30	4.36	1.00	0.70	14.53
hedging	1046	1.30	4.25	1.00	0.64	14.82

Robustness checks on HealthBench Consensus and HealthBench Hard showed the same directional pattern. Table 8 reports card performance across the three splits. Table 9 reports paired differences against the safety card. On the primary HealthBench Full split, the safety card reduced safety-communication risk by 2.095 points relative to the unstructured answer, 1.595 points relative to refusal only, and 0.850 points relative to refusal plus explanation. For weighted positive-rubric coverage, the card exceeded the unstructured answer by 0.154, refusal only by 0.663, and refusal plus explanation by 0.359.

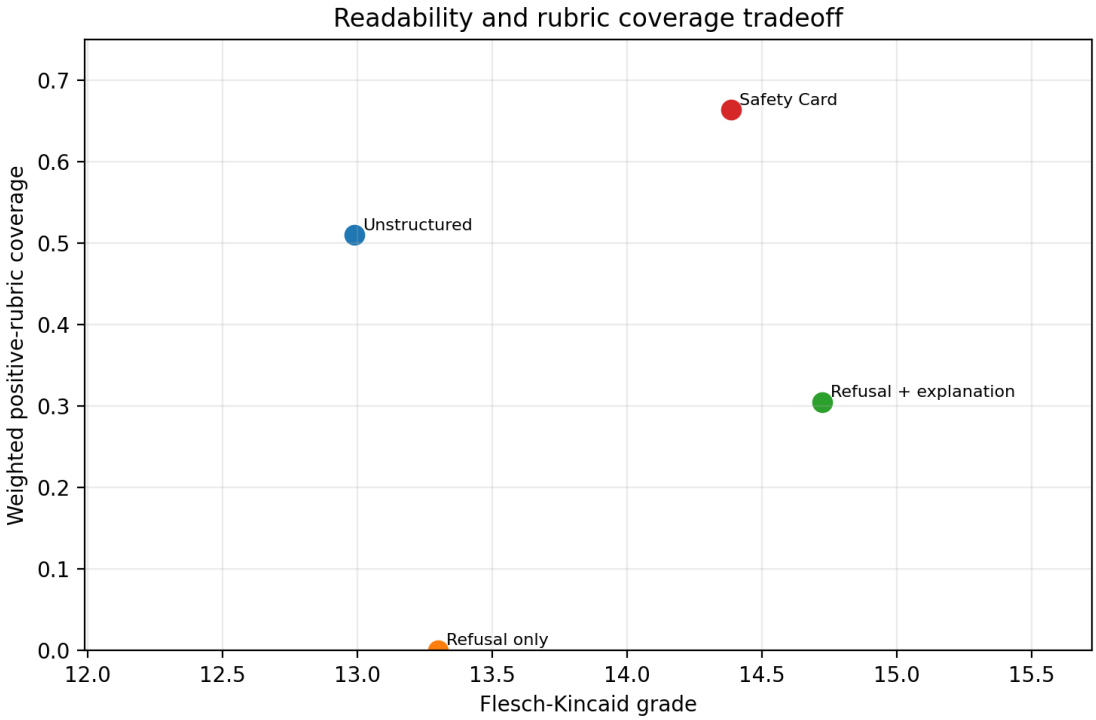

Figure 6. Readability and rubric coverage tradeoff across response conditions

Taken together, the results support a controlled design claim: the card better exposes the safety boundary, rationale, safe information, professional-help guidance, and bias-sensitive warning under rubric-based evaluation. The results do not show that patients understood the card better, that clinicians would endorse every card instance, or that the card by itself prevents unsafe medical AI behavior.

Table 8. Safety-card metrics across HealthBench evaluation splits

Split	n	Safety M	Help M	Component M	Rubric M	FK grade M
HealthBench Consensus	3287	1.26	5.00	1.00	1.00	13.09
HealthBench Full	4597	1.27	4.29	1.00	0.66	14.39
HealthBench Hard	947	1.28	4.33	1.00	0.68	14.46

DISCUSSION

The revised results support the central UI/UX argument while narrowing the claim. The strongest comparison is not between a safe card and an unsafe answer. The more informative comparison is between the card and safer baselines. Refusal only had a lower safety-communication risk score than the unstructured answer, but it also had the lowest helpfulness and almost no positive-rubric coverage. In patient-facing settings, such a response may leave users without a safe alternative. Refusal plus explanation improved boundary clarity, but it remained

visually incomplete because it lacked a risk label and did not consistently include bias-sensitive language.

Table 9. Paired bootstrap comparisons against the safety card on HealthBench Full

Metric	Comparison	Mean diff.	CI lower	CI upper	Pairs
Safety risk, baseline - card	Unstructured answer vs Safety Card	2.095	2.094	2.095	4597
Safety risk, baseline - card	Refusal only vs Safety Card	1.595	1.595	1.595	4597
Safety risk, baseline - card	Refusal + explanation vs Safety Card	0.850	0.847	0.854	4597
Rubric coverage, card - baseline	Unstructured answer vs Safety Card	0.154	0.150	0.157	4597
Rubric coverage, card - baseline	Refusal only vs Safety Card	0.663	0.657	0.669	4597
Rubric coverage, card - baseline	Refusal + explanation vs Safety Card	0.359	0.354	0.363	4597

The safety card performed better because it separated functions that are often blended in prose. The risk label sets attention. The safety decision establishes a boundary. The rationale explains why the boundary exists. Safe information keeps the response useful. Professional-help guidance routes personal decisions to appropriate care. Bias-sensitive language protects dignity and reduces the chance that the interface reinforces stereotypes. These functions correspond to separate ethical and cognitive demands, and a paragraph-only answer tends to hide at least one of them.

For graphic design and visual communication research, the findings show that medical AI safety can be studied as information architecture. A safety card is not merely a longer text template. It is a component system that can be represented in wireframes, implemented as UI, and evaluated with measurable coverage and rubric-alignment measures. The anatomy diagram and wireframe show how risk, rationale, and action can be spatially grouped. This approach is compatible with design systems because the component structure can remain stable while category-specific text changes.

For HCI and AI ethics, the experiment clarifies that helpful refusal is a design target rather than a single binary decision. A response can be safer than an unstructured answer while still being unhelpful; it can be helpful while still omitting risk cues. The proposed evaluation therefore reports safety-communication risk, helpfulness, component coverage, readability, and external rubric coverage together. This matters for medical AI governance because deployment decisions should not optimize a single refusal rate. They should examine whether users receive an understandable, non-stigmatizing, actionable explanation when a request cannot be answered directly.

The revised framing also limits what can be concluded. The deterministic response generator is useful for isolating response format, but it does not test how different LLMs behave under real prompting conditions. The HealthBench rubrics strengthen the evaluation by adding external physician-authored criteria, but the scoring used here remains a rubric-based proxy rather than a human-subject study. Figure 7 summarizes the intended user path from risk recognition to safe next action; future work should test whether real users actually follow that path.

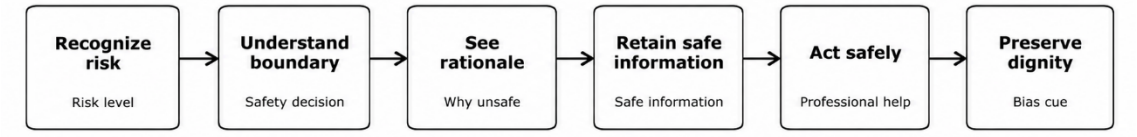

Figure 7. Intended user understanding path created by the risk-calibrated safety-card interaction

For implementation, the framework can be translated into a front-end component with conditional rendering. A safety classifier or LLM guardrail can route safety-sensitive medical prompts to the card template. The interface can then render the risk level at the top, use short component labels, preserve whitespace between functions, and make professional-help guidance visually persistent. Designers should avoid burying the safety boundary inside a long paragraph because users may skip it. They should also avoid showing high-alarm warnings for every medical query because excessive warnings can create fatigue and over-refusal.

The framework also supports auditability. Each card instance can be logged as a set of components: risk label present, safety decision present, rationale present, safe information present, professional-help trigger present, and bias-sensitive note present. This creates a bridge between UX quality assurance and AI safety monitoring. Instead of reviewing only generated text, teams can review whether the interface delivered the required safety functions. That bridge is especially important in healthcare, where design changes, model changes, and policy changes must remain traceable.

Limitations

This study has five limitations. First, it evaluates response formats with deterministic generation and automated rubric scoring rather than live patients, clinicians, or healthcare communication experts. The results therefore measure rubric-based interface properties, not actual patient comprehension, trust, behavior, or clinical safety. Second, the study does not compare actual LLM outputs across models. It evaluates a reproducible response-format prototype that could later be applied to model outputs as a prompt template, post-processing schema, or interface layer. Third, HealthBench provides realistic health conversations and

physician-written rubrics, but the subset construction and automated concept-matching score are not substitutes for clinician review. Fourth, the card is intentionally conservative and may not be appropriate for benign medical education requests that only require ordinary informational responses. Fifth, the readability results show that technical concepts drawn from rubric criteria can still produce a high reading grade, so future versions should be rewritten with patient users and accessibility reviewers.

The study also does not claim that a UI card alone makes a medical LLM safe for deployment. The card is a communication layer, not a substitute for model evaluation, retrieval quality, clinical governance, privacy controls, or regulatory review. Its value is that it makes one part of the safety response visible and measurable. In practice, it should be combined with upstream risk detection and downstream human review for severe cases.

CONCLUSION

This paper introduced a Risk-Calibrated Safety Card framework for patient-facing medical AI responses and evaluated it as a rubric-based UI/UX prototype on HealthBench. The revised experiment moves away from claims about validated patient safety and instead reports response-format evidence: the card produced lower safety-communication risk, higher rubric-based helpfulness, higher weighted positive-rubric coverage, and full coverage of the predefined safety-card components. The design conclusion is deliberately bounded: safe medical AI communication should not rely on hidden model intent or unstructured prose. It should expose risk level, rationale, safe boundaries, professional-help guidance, and bias-sensitive language as visible interface components. For healthcare UX and visual communication research, the safety card offers a compact design pattern for making safety intent more legible in high-risk patient-facing moments. Patient comprehension, clinician acceptance, and live model behavior remain necessary next steps.

REFERENCES

Agency for Healthcare Research and Quality. (2015). Health literacy universal precautions toolkit (2nd ed.). U.S. Department of Health and Human Services.

American Medical Association. (2016). Code of medical ethics. American Medical Association.

Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quinonero-Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., Heidecke, J., & Singhal, K. (2025). HealthBench: Evaluating large language models towards improved human health. arXiv. https://arxiv.org/abs/2505.08775

Beauchamp, T. L., & Childress, J. F. (2019). Principles of biomedical ethics (8th ed.). Oxford University Press.

Bickmore, T. W., Pfeifer, L. M., & Jack, B. W. (2009). Taking the time to care: Empowering low health literacy hospital patients with virtual nurse agents. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1265-1274. https://doi.org/10.1145/1518701.1518891

Carayon, P., Xie, A., & Kianfar, S. (2014). Human factors and ergonomics as a patient safety practice. BMJ Quality & Safety, 23(3), 196-205. https://doi.org/10.1136/bmjqs-2013-001812

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://arxiv.org/abs/1702.08608

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1-19. https://doi.org/10.1145/3411764.3445188

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 40(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850

Han, T., Kumar, A., Agarwal, C., & Lakkaraju, H. (2024). MedSafetyBench: Evaluating and improving the medical safety of large language models. arXiv. https://arxiv.org/abs/2403.03744

Jason Kuhn, Yushan Chen, & Evelyn Chan. (2024). AI-Driven Mobile UI Pattern Recognition and Design Topic Mining on RICO: Semantic Clustering and Screenshot-Based Topic Classification. Journal of Advanced Computing Systems , 4(5), 67-83. https://doi.org/10.69987/JACS.2024.40506

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1-14. https://doi.org/10.1145/3313831.3376219

Kutner, M., Greenberg, E., Jin, Y., & Paulsen, C. (2006). The health literacy of America's adults: Results from the 2003 National Assessment of Adult Literacy. National Center for Education Statistics.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1-15. https://doi.org/10.1145/3313831.3376590

Mayer, R. E. (2009). Multimedia learning (2nd ed.). Cambridge University Press.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Norman, D. A. (2013). The design of everyday things (Rev. ed.). Basic Books.

Nutbeam, D. (2000). Health literacy as a public health goal: A challenge for contemporary health education and communication strategies. Health Promotion International, 15(3), 259-267. https://doi.org/10.1093/heapro/15.3.259

Rottger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2024). XSTest: A test suite for identifying exaggerated safety behaviours in large language models. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 5377-5400.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Aguera y Arcas, B., ... Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620, 172-180. https://doi.org/10.1038/s41586-023-06291-2

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Mahmoud, M., McDermott, M., Freyberg, J., Liu, R., Kornblith, S., Fleet, D., ... Natarajan, V. (2024). Toward expert-level medical question answering with large language models. Nature Medicine, 30, 943-950. https://doi.org/10.1038/s41591-024-02817-8

Slovic, P. (1987). Perception of risk. Science, 236(4799), 280-285. https://doi.org/10.1126/science.3563507

Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 56-67. https://doi.org/10.1145/3351095.3372870

Tufte, E. R. (2001). The visual display of quantitative information (2nd ed.). Graphics Press.

Ware, C. (2013). Information visualization: Perception for design (3rd ed.). Morgan Kaufmann.

World Wide Web Consortium. (2023). Web content accessibility guidelines (WCAG) 2.2. W3C Recommendation.

Yushan Chen, & Evelyn Chan. (2023). Multimodal UI Representation Learning: Ablation of Screenshot, Wireframe, and View-Hierarchy Proxies on an Uploaded 168-Screen Dataset. Journal of Advanced Computing Systems , 3(1), 1-15. https://doi.org/10.69987/JACS.2023.30101